Análisis de Métricas de Similaridad Usadas en un Algoritmo de Filtro Colaborativo Basado en el Usuario Para Recomendar Materias de Pregrado

Del Pino, J.; Salazar, G.; Cedeño, V. Msc.
Facultad de Ingeniería en Electricidad y Computación
Escuela Superior Politécnica del Litoral (ESPOL)
Campus Gustavo Galindo, Km 30.5 vía Perimetral
Apartado 09-01-5863. Guayaquil-Ecuador
{jdelpino,gsalazar,vcedeno}@fiec.espol.edu.ec

Resumen

En la actualidad, los sistemas de recomendación son ampliamente utilizados debido a su capacidad de analizar preferencias de usuarios para sugerirles items afines. En el ámbito educativo, el momento en el cual una materia es tomada por un estudiante a lo largo de su carrera representa un factor primordial para realizar recomendaciones de materias a tomar cada semestre .El presente trabajo representa un esfuerzo por adaptar un recomendador de filtro colaborativo basado en el usuario para el diseño e implementación de un recomendador de materias a estudiantes de pregrado, tomando en cuenta la historia académica de los mismos, además de determinar una métrica de similaridad idónea para realizar las recomendaciones. El sistema propuesto puede ser adaptado a cualquiera de las carreras de Espol y servir de base para futuras implementaciones. Los resultados obtenidos en las pruebas muestran un acierto del ochenta por ciento de las materias recomendadas respecto a las que los estudiantes tomaron en realidad.

Palabras Claves: *Sistema de recomendación, Recomendador de materias, Filtro colaborativo, Guía académica.*

Abstract

Actually recommender systems are broadly used thanks to their potential on analyzing user preferences and suggesting items to them. In the academic scope, the moment a subject is taken by a student throughout his career represents an important factor when recommending subjects each semester. The present work represents an effort to adapt a collaborative filter and user-based recommender system to design and implement a subject recommender for undergraduate students considering their academic history, and also determine the ideal similarity metric to perform recommendations. The proposed system could be adapted to any Espol career and provides a basis for future implementations. The results show eighty percent of success of recommended subjects over subjects taken by students.

Keywords: *Recommender system, Subject recommender, Collaborative Filter, Academic guidance.*

.

1. Introducción

Durante los últimos años, los sistemas de recomendación han sido herramientas que han contribuido a mejorar la experiencia de los usuarios en las aplicaciones que estos utilizan. Numerosas empresas han apostado por el uso de estas herramientas para lograr que sus sitios web posean un ambiente personalizado, en donde el contenido que se les presenta a sus usuarios está estrechamente atado a sus preferencias personales. Algunas empresas que usan recomendadores en sus sitios web son Amazon, Netflix, Facebook y Twitter; los cuales sugieren compras de productos, películas, amistades y personas a seguir, respectivamente. La gran cantidad de información filtrada de algunos usuarios sirve para ayudar a otros, resultando en un entorno colaborativo. La aplicación de los sistemas de recomendación en el plano comercial es muy conocido, sin embargo su uso en otras áreas como la académica ha despertado el interés de investigadores que ven en las instituciones educativas una gran fuente de información apta para realizar recomendaciones. La precisión de los sistemas de recomendación basados en filtro colaborativo, específicamente los basados en el usuario en gran parte

se debe al cálculo que mide la similaridad entre usuarios, por ello se realiza un análisis de dos métricas de similaridad: *Tanimoto* y *LogLikelihood*.

El presente artículo introduce una solución informática que permite adaptar un recomendador basado en el usuario para realizar recomendaciones cuando los estudiantes requieran sugerencias de registros de materias, las mismas que estarán basadas en experiencias pasadas de otros estudiantes de acuerdo a su historial académico.

El contenido de este artículo se distribuye de la siguiente manera: en la sección 2 se discute la motivación que llevó a escribir el artículo. En la sección 3 se describe el trabajo relacionado, seguida de la sección 4 dedicada a definir conceptos básicos sobre sistemas de recomendaciones que usan filtro colaborativo y son basados en el usuario. En la sección 5 se describe la metodología empleada para el desarrollo de la solución. La sección 6 incluye el análisis de los resultados. Finalmente, en la sección 7 concluimos y detallamos el trabajo futuro.

2. Motivación

Dada la gran cantidad de información que posee Espol sobre los registros de materias aprobadas de sus estudiantes resulta provechoso procesarla aplicando una tecnología innovadora como los sistemas de recomendación y obtener beneficios para los estudiantes. De la misma manera, podemos aportar al desarrollo tecnológico de los servicios informáticos de la universidad e intentar resolver el problema de la dificultad que tienen los estudiantes para decidir qué materias tomar en un nuevo semestre.

3. Trabajos relacionados

Una recomendación personalizada requiere de un conocimiento amplio de la información académica del estudiante por parte de los asesores académicos, por ello se han implementado sistemas que pretenden cumplir las mismas funciones de las personas que realizan dicho trabajo. Haciendo las veces de un asesor académico, AARCON [1] es un sistema de recomendación basado en casos [2] que recomienda los cursos que debe tomar un estudiante de la Universidad De Paul basándose en el programa académico de la institución, el historial de cursos tomados por el estudiante y la información de otros estudiantes con un historial similar.

Una función semejante cumple RARE [3], que recomienda cursos de maestría basándose en reglas de asociación. Además de cumplir funciones de asesores académicos algunos sistemas usan información directa de los mismos asesores junto con evaluaciones de los estudiantes sobre los cursos que han aprobado, como Course Agent [4]. Otros sistemas combinan distintas técnicas de recomendación, por ejemplo usan filtros colaborativos junto a un sistema de recomendación basado en contenido.

También se han usado filtros colaborativos para recomendar materias de especialización y electivas en el sistema de educación español de escuelas secundarias [5].

4. Conceptos y Tecnologías Empleadas

Un sistema de recomendación de filtro colaborativo se basa en la información que se obtiene de la interacción de usuarios con ítems, por ejemplo la compra de un artículo en línea. La interacción entre un usuario y un ítem es vista como una preferencia que tiene el primero por el segundo. En algunos casos la interacción tiene un valor de preferencia que la representa y determina el grado de preferencia del usuario por el ítem. El conjunto de usuarios y sus preferencias forman un modelo de datos de entrada para un sistema de recomendación.

Las preferencias de algunos usuarios sirven para recomendar ítems a otros usuarios que tienen preferencias similares. Un grupo de usuarios que guardan similaridad entre sí es llamado *vecindad*. Una vecindad puede estar restringida de dos maneras; por el número máximo de usuarios que pueden conformarla, llamada vecindad "N Cercanos"; o por un valor mínimo de similaridad que debe haber entre los usuarios, llamada vecindad "Umbral" [6].

Para determinar en qué medida son similares los usuarios se utilizan métricas de similaridad. Las métricas pueden usarse con o sin valores de preferencias. En el presente trabajo se usaron dos métricas de similaridad que no usan valores de preferencia, *Tanimoto* y *Loglikelihood* [7][8].

Para recomendar items a un usuario se escogen aquellos items que los demás usuarios de su vecindad han preferido pero el usuario al que daremos la recomendación aún no conoce. La relevancia de los items recomendados se determina mediante cálculos que estimen el grado de preferencia que el usuario al que se le hace la recomendación tendría por cada uno de esos ítems [9].

La lógica descrita previamente corresponde al método de recomendación tradicional usando filtros colaborativos basados en el usuario. Esta lógica puede ser aplicada al escenario académico, en el que los usuarios son los estudiantes, los ítems son las materias y la interacción entre ambos se da cuando el estudiante aprueba dicha materia, por lo tanto a un estudiante se le recomiendan materias que él aún no ha aprobado pero otros estudiantes similares a él sí.

Un sistema de recomendación de materias será más efectivo mientras más de las materias recomendadas a un estudiante antes de empezar un nuevo semestre, estén entre las materias en las que él decide registrarse.

Al aplicar la lógica por defecto de los filtros colaborativos en la recomendación de materias, varios estudiantes que van a registrarse en un mismo semestre y han aprobado las mismas materias recibirían las mismas recomendaciones. Sin embargo debido a la variedad de opciones y libertad de decisión que tienen los estudiantes, no todos ellos se registrarán en las materias recomendadas, por lo tanto el recomendador no sería efectivo. Es necesario tomar en cuenta otros factores además de las materias aprobadas para hallar un patrón que permita predecir el comportamiento de un estudiante al registrarse en un nuevo semestre. Uno de esos factores es el orden en que los estudiantes aprueban las materias, dado por su historial académico.

5. Materiales y Método

5.1. Adaptación del Recomendador

El método propuesto para la adaptación del recomendador basado en el usuario consiste en introducir un nuevo concepto de vecindad. Esta vecindad, que se ha definido como *Vecindad por Historial,* contiene el conjunto de estudiantes que mantienen la mayor semejanza en el orden de aprobación de materias con respecto al estudiante que recibe la recomendación, al cual lo llamaremos de ahora en adelante como *Estudiante E*. Esta vecindad es el producto de un recorrido en el historial académico del *Estudiante E*. Para recomendar materias en un Semestre *N*, se calculan las vecindades individuales para cada uno de los semestres anteriores a *N*, es decir, desde el primer semestre, hasta el semestre *N-1*. Por cada Semestre *Si*, se establece una Vecindad *Vi*, la lógica del método tradicional. Cada Vecindad *Vi* contiene un conjunto de estudiantes que mantienen un grado de similaridad con el *Estudiante E* en un *Semestre Si*, con respecto a las materias tomadas en ese semestre.

Con el conjunto de Vecindades *Vi*, se construye una *Tabla de Frecuencia (Hashmap)* cuya estructura contiene: los IDs de los estudiantes (*Key*), y un valor de repetición (*Value)*. El objetivo de la *Tabla de Frecuencia* es registrar el número de semestres en los que cada uno de los estudiantes evaluados mantuvo una similaridad con el *Estudiante E*. Dicha similaridad está determinada por la métrica de similaridad utilizada en el cálculo. En otras palabras, la *Tabla de Frecuencia* puede ser vista como una lista ordenada en forma descendente basada en el puntaje de aparición de un estudiantes en las vecindades *Vi*.

Una vez obtenida la *Tabla de Frecuencia*, es posible obtener la *Vecindad por Historial.* Para su efecto, se utilizan diversos métodos. En nuestro trabajo, hemos considerado dos: *Umbral* y *N Cercanos*. De ellos depende la cantidad de Estudiantes filtrados a la *Vecindad por Historial*. Paralelamente a la construcción de la *Vecindad por Historial*, se calcula la. *Similaridad Final* de cada uno de los estudiantes pertenecientes a la *Tabla de Frecuencia*.

La *Similaridad Final* es el promedio de todos los valores de similaridad que un estudiante registró en los *Semestres Si* con respecto al *Estudiante E.*

La *Similaridad Final* es usada en el proceso de *estimación de preferencia* de una materia, de esta manera las materias que los estudiantes de la *Vecindad por Historial* aprobaron en su *Semestre N*, son los cursos que se entregan como recomendación al *Estudiante E*.

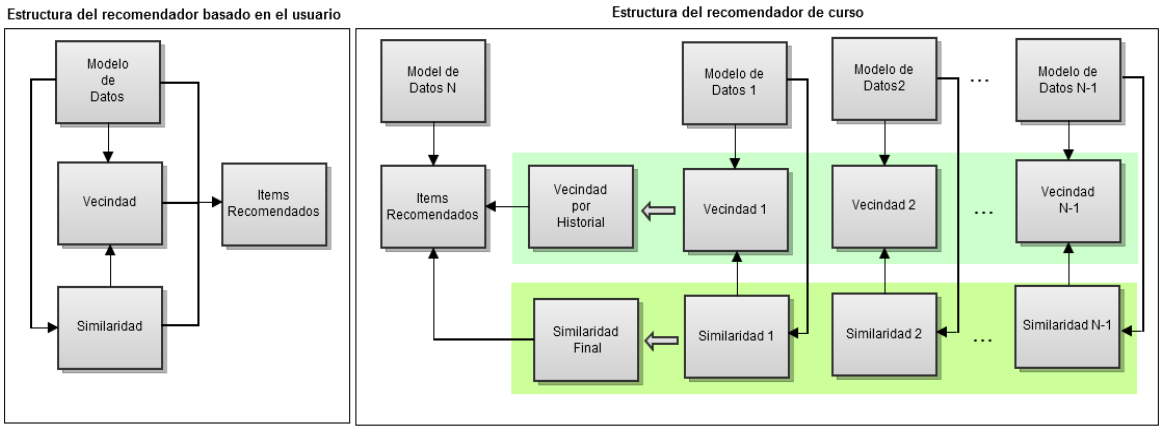
La Figura 1 muestra las diferencias entre el método de recomendación tradicional de filtro colaborativo basado en el usuario, al método modificado y propuesto en el presente trabajo (derecha). Como se puede observar en la gráfica, usando el método tradicional se realiza la recomendación de ítems basado en un *Modelo de Datos (Datamodel)* único, para posteriormente obtener una única Vecindad luego de establecer Similaridades. En contraste, en el trabajo propuesto, los ítems a recomendar provienen del *Modelo de Datos* que representa el *Semestre N*, mientras que la vecindad y la similaridad para realizar la recomendación provienen de múltiples vecindades, creadas a partir de múltiples similaridades y sus *Modelos de Datos*.

Sin la modificación del algoritmo, propuesta en el presente trabajo, un estudiante no podría tener una recomendación por semestre. En su lugar, tendría una recomendación de todo el universo de materias de la carrera, basado en las materias que ha tomado todo el universo de Estudiantes.

La adaptación del recomendador se logró modificando los algoritmos de filtro colaborativo de Mahout, una librería de código abierto que implementa algoritmos auto aprendizaje e inteligencia colectiva [10].

5.2. Pruebas realizadas

Usando el sistema de recomendaciones implementado se recomendaron 6 materias por semestre a los estudiantes de la carrera de Ingeniería en Telecomunicaciones de la FIEC (Facultad de Ingeniería Eléctrica y Computación), ESPOL. Se usó esta carrera por ser la que posee el mayor número de estudiantes registrados en la FIEC, esto es fundamental para el recomendador, ya que la calidad de las recomendaciones está en gran parte determinada por la cantidad y la calidad de los datos usados para realizarlas [10]. Además, la ausencia de especializaciones en la carrera evita que el grupo de estudiantes registrados en ella se dividan en subgrupos, evadiendo así validaciones complejas que introduzcan cambios que pueden afectar los resultados finales de las recomendaciones. La información utilizada como entrada del sistema son los registros de



materias aprobadas del grupo de estudiantes

Figura 1. Elementos del recomendador basado en el usuario (izquierda). Elementos del recomendador de curso basado en historial académico (derecha).

mencionado que han ingresado a la carrera desde el primer término del año 2007 hasta el primer término del año 2010, período en el que hay 8 semestres por lo que las predicciones de materias se realizan desde el segundo hasta máximo el octavo semestre de cada estudiante. Para el primer semestre no se realizan recomendaciones usando la solución propuesta porque no se conoce historial académico previo a ese semestre.

Para determinar qué tan confiable y efectivo es el sistema de recomendaciones tomamos las materias que un estudiante aprobó en un semestre y las comparamos con las que el sistema predice para ese semestre, es decir comparamos datos reales con datos predichos. Para efecto de describir las pruebas realizadas nos referimos como materias aprobadas a aquellas materias que el estudiante aprobó en el semestre del que se piden recomendaciones. Mientras más materias aprobadas aparecen entre las recomendaciones para un estudiante, más efectivo es el recomendador.

La comparación entre materias predichas y aprobadas se realiza usando una métrica de recuperación de información llamada *Recall*, que es el porcentaje de materias aprobadas que fueron recomendadas. Para realizar las recomendaciones por semestre y por estudiante se usaron cuatro tipos de configuraciones en el recomendador. Dos variaciones de métrica de similaridad: Tanimoto, Loglikelihood; y dos de filtrado de vecindad: N Cercanos, Umbral. La manera en la cual se graficaron los resultados fue mediante la variación de los valores de Umbral y N Cercanos. Los valores de Umbral a usarse en las pruebas son: 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, mientras que los valores de tamaño de vecindad para N Cercanos son: 1, 2, 4, 8, 16, 32, 64, 128. Se calculó el promedio de Recall de todos los estudiantes para cada semestre desde el primer término del año 2007 y variando en cada uno de ellos la configuración de métrica de similaridad y filtrado de vecindad.

6. Análisis de Resultados

En la Figura 2 se muestran los promedios de *Recall* para los diferentes tamaños de vecindad N Cercanos para recomendaciones hechas en el semestre 5. El eje vertical representa el promedio de *Recall* del semestre, mientras que el eje horizontal representa el tamaño de la vecindad. Para los demás semestres la tendencia de la curva fue similar. Para las dos métricas de similaridad: *Loglikelihood* y *Tanimoto*, se incrementa el valor promedio de *Recall* conforme se va incrementando el tamaño de la vecindad, esto se debe a que las recomendaciones que usan un tamaño de vecindad pequeño se basan en un número de estudiantes tan reducido, que no se cuenta con una cantidad suficiente de ítems a analizar para la recomendación. Luego la curva crece de manera exponencial conforme la vecindad se agranda. Al pertenecer más estudiantes a la *vecindad por historial*, las materias recomendadas no solo se escogerán por haber sido aprobadas por los estudiantes con un grado de similaridad muy cercano al estudiante objetivo, sino que también se verán influenciadas por la tendencia de la mayoría de los estudiantes al aprobar la materia en el semestre, ya que mientras más estudiantes aprueban la materia el valor de preferencia de ésta va a aumentar, inclusive si los estudiantes que la aprobaron tuvieron un valor de similaridad normal o bajo.

Los valores óptimos de promedio de Recall para todos los semestres se alcanzaron con el tamaño de vecindad 64 y la métrica de similaridad *Loglikelihood*. En promedio el valor de *Recall* para todos los semestres fue de 0.83. Cabe destacar que la métrica de similaridad *Tanimoto* con un tamaño de vecindad de 32, alcanza valores cercanos a *Loglikelihood* con tamaño 64 para todos los semestres. El promedio de *Recall* alcanzado por

Tanimoto es 0.81, sin embargo para las comparaciones con los resultados de tipo de vecindad *Umbral* se usará *Loglikelihood* por tener un *Recall* superior.

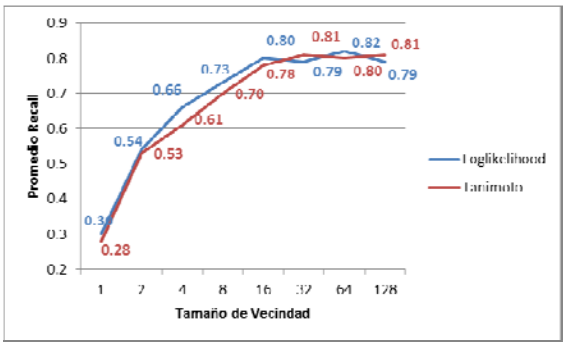


Figura 2. Promedio recall obtenido para los diferentes tamaños de vecindad en el semestre 5 para las métricas Loglikelihood y Tanimoto.

En la Figura 3 se muestran los promedios de *Recall* para los diferentes valores de umbral para las vecindades de tipo *Umbral* para recomendaciones hechas en el semestre 5. El eje vertical representa el promedio de Recall del semestre, mientras que el eje horizontal representa el valor de umbral. El valor de promedio de *Recall* decrece conforme se va aumentando el valor de umbral, esto se debe a que el incremento del valor del umbral implica que los usuarios de la vecindad por historial deben tener un historial académico muy similar al del estudiante objetivo, es decir que debieron haber aprobado las materias de igual manera para un número mayor de semestres, lo que provoca que el número de estudiantes en la vecindad por historial sea cada vez menor. Para los demás semestres el comportamiento de la curva es similar pero a medida que avanza el número de semestre las curvas empiezan a decrecer y dejan de ser rectas como al inicio. Esto se debe a que a medida que avanzamos en el flujo de una carrera se incrementa el número de caminos y formas de orden en las que se pueden aprobar las materias, teniendo así grupos cada vez más pequeños de usuarios que han recorrido su flujo de la misma manera.

Esto ocasiona que la mayoría de estudiantes en la tabla de frecuencia tenga valor de frecuencia cada vez menores, y al realizarse el filtro por umbral las vecindades contendrán menos estudiantes conforme aumente el semestre y se aumente el valor de umbral. Como se mencionó anteriormente, las recomendaciones basadas en vecindades pequeñas habitualmente no alcanzan valores de acierto altos. Para el tipo de vecindad *Umbral* los mejores promedios de *Recall* para la mayoría de semestres se obtuvieron con *Loglikelihood* con valor de umbral 0.4 y *Tanimoto* con valor 0.3, ambos tuvieron un promedio de *Recall* de 0.79 para todos los semestres. Sin embargo *Tanimoto* posee diferencias de valor considerables entre sus valores máximos y mínimos obtenidos en cada semestre, a diferencia de *Loglikelihood* que mantiene valores estables para todos los semestres. Por este motivo para las comparaciones con los valores de las vecindades de tipo *N Cercanos* se usó a *Loglikelihood* con valor de umbral 0.4 como la combinación óptima.

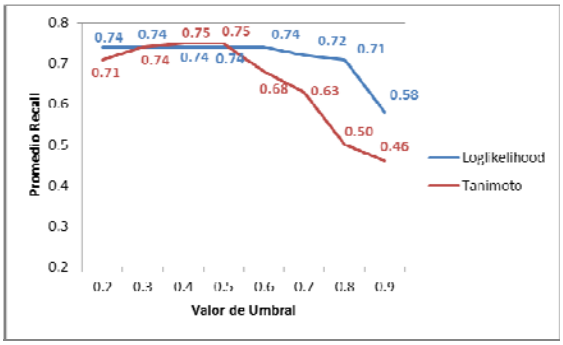


Figura 3. Promedio de Recall obtenido para los diferentes valores de umbral en el semestre 5 para las métricas Loglikelihood y Tanimoto.

En la Figura 4 y 5 se grafican los valores de Recall obtenidos en todos los semestres para las dos combinaciones seleccionadas: Loglikelihood valor 64 para N Cercanos, Loglikelihood valor 0.4 para Umbral. Con el tipo de vecindad N Cercanos los valores de acierto en las recomendaciones fueron similares para todos los semestres, manteniendo una tendencia más estable que usando el tipo de vecindad Umbral, en donde hubo una mayor variación en los valores de acierto. Se determinó que Loglikelihood con una vecindad de tipo N Cercanos de tamaño 64 es la mejor combinación para el recomendador de materias, con un promedio de acierto de 0.83 en las recomendaciones respecto a las materias que los estudiantes aprobaron en la realidad, este valor de acierto se lo obtuvo calculando un promedio de los valores de Recall de todos los semestres.

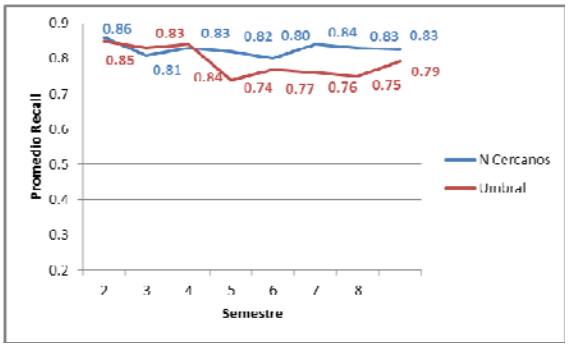


Figura 4. Resultados promedio de Recall usando métrica de similaridad Loglikelihood con los tipos de vecindad N Cercanos y Umbral del segundo al octavo semestre.

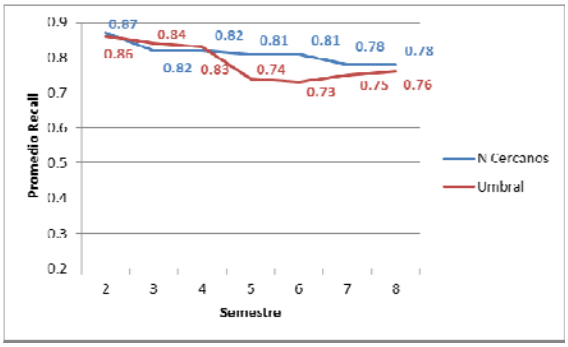


Figura 5. Resultados promedio de Recall usando métrica de similaridad Tanimoto con los tipos de vecindad N Cercanos y Umbral del segundo al octavo semestre.

7. Conclusiones

En conclusión la adaptación del recomendador se realizó de manera exitosa. Se consiguió un valor de acierto de 0.83 en las recomendaciones con respecto a las materias que el estudiante decidió tomar en la realidad. Este valor se lo consiguió usando la métrica de similaridad Loglikelihood y en tipo de vecindad N Cercanos con un tamaño igual 64. Los estudiantes que pertenecen a una vecindad de este tamaño para la métrica mencionada tienen en promedio un valor de similaridad de 0.57 con una desviación estándar de 0.11. Sesenta y cuatro representa un tamaño de vecindad óptimo dentro de los valores asignados en las pruebas. Futuros estudios podrán determinar el valor óptimo del tamaño de vecindad para el recomendador.

La métrica de similaridad Loglikelihood demostró darle un comportamiento más certero y uniforme al recomendador en cuanto a sus predicciones. En contraparte Tanimoto alcanzo altos valores en algunas de las pruebas realizadas, pero el recomendador en general no siempre entregó recomendaciones altamente acertadas en todos los semestres.

Con respecto al tipo de vecindad, el recomendador no sostuvo buenos valores con el uso de vecindad Umbral, al aumentar el valor de umbral el grado de similaridad es muy estricto, por lo que no hay muchos estudiantes en los cuales basar la comparación.

El éxito del recomendador de materias radica en encontrar patrones de grupos. El filtrado de vecindad por medio de N Cercanos con tamaño de vecindad 64 le permitió al recomendador alcanzar los valores de acierto más altos en las pruebas y determinar para cada estudiante un patrón que permita predecir con mayor precisión su comportamiento de registros de materias.

Con un tamaño de vecindad excesivamente grande ingresaran estudiantes que introduzcan “ruido” en las recomendaciones, ya que con estas grandes vecindades el recomendador empieza a entregar como recomendación las materias que la mayoría de estudiantes aprobó en el semestre, sin mayor distinción.

El tiempo de recomendación a un estudiante fue de aproximadamente 10 segundos, con un hardware de 2GB de memoria RAM y un procesador de 1.2 Ghz. Por lo que las recomendaciones deben ser calculadas de manera previa a las consultas, ya que alcanzar buenos tiempos de respuesta en tiempo real implica altos costos en hardware.

Las recomendaciones no necesariamente son las ideales, ya se debe considerar que las decisiones que tomaron los estudiantes en los que se basa la recomendación, no necesariamente involucra una combinación eficiente en la manera de aprobar el flujo de la carrera. Se debe considerar añadir al proceso de recomendación otros factores de criterio de elección de materias que permitan hallar patrones más precisos.

10. Referencias

[1] O' Mahonny, Michael P. , Smith, Barry. , A Recommender System for Online Course Enrolment, Proceedings of the 2007 ACM conference on Recommender systems, 2007, pp. 133-136.

[2] Smyth, B., Case-based recommendation, The adaptive web, Springer, 2007, pp. 342-376.

[3] Bendakir N. , Aimeur E., Using association rules for course recommendation, In Proceedings of the AAAI Workshop on Educational Data Mining, 2006, pp. 31–40.

[4] Farzan R, Brusilovsky P., Social navigation support in a course recommender system. In Proceedings of the 4th International Conference on Adaptive Hypermedia and Apadtive Web-based Systems, 2006, pp. 91–100.

[5] Castellano E.J., Martínez L., A web-decision support system based on collaborative Filtering for academic orientation. case study of the spanish secondary school, Journal of Universal Computer Science, 2009.

[6] Breese John S., Heckerman David, Kadie Carl, Empirical Analysis of Predictive Algorithms for Collaborative Filtering, Microsoft Research,Microsoft Corporation, 1988. .

[7] J.M., The probabilistic basis of Jaccard's index of similarity. Systematic Biology, 2006, pp. 380-385.

[8] Dunning, T. Surprise and Coincidence, http://tdunning.blogspot.com/2008/03/surprise-and-coincidence.html, 2008.

[9] Adomavicius, G., Tuzhilin, A. Toward the next generation of recommender systems: A survey of the stateof-the-art and possible extensions, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, IEEE Computer Society, 2005.

[10] Owen S., Anil R., Dunning T., Friedman E., Mahout in Action, Manning Publications Co., 2009, pp. 4-65